PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2001-344259
(43)Date of publication of application : 14.12.2001

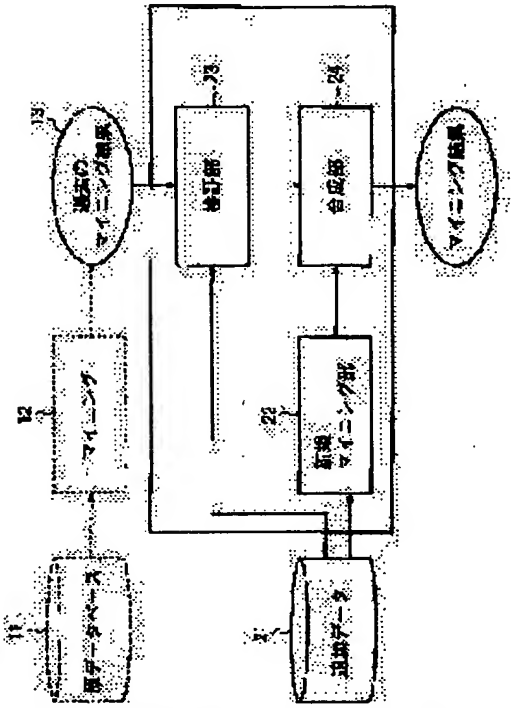
(51)Int.CI. G06F 17/30
G06F 12/00

(21)Application number : 2000-162080
(22)Date of filing : 31.05.2000

(71)Applicant : TOSHIBA CORP
(72)Inventor : KOYANAGI SHIGERU
SAKAI HIROSHI
NAKASE AKIHIKO
KUBOTA KAZUTO

(54) METHOD AND DEVICE FOR INFORMATION ANALYSIS

(57)Abstract:
PROBLEM TO BE SOLVED: To provide an incremental mining method which increases the data mining speed at the time of data addition or deletion.
SOLUTION: In the correlation rule discovery as a data mining technique, past mining results are preserved, and when data are added or deleted, a past database is not accessed and the result obtained by verifying the past mining results with respect to added data and the mining result related to added data are synthesized to perform mining of entire data.



LEGAL STATUS

[Date of request for examination] 31.05.2000
[Date of sending the examiner's decision of rejection]
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]
[Date of final disposal for application]
[Patent number]
[Date of registration]
[Number of appeal against examiner's decision of rejection]
[Date of requesting appeal against examiner's decision of rejection]
[Date of extinction of right]

* NOTICES *

1. This document has been translated by computer.So the translation may not reflect the original precisely.
2. **** shows the word which can not be translated.
3.In the drawings, any words are not translated.

CLAIMS

[Claim(s)]
[Claim 1] The information-analysis method characterized by providing the following. The step which is the information-analysis method using the correlation rule discovery technique, analyzes the aforementioned additional information and acquires the 2nd analysis result information while verifying the existing analysis result information for the aforementioned additional information and acquiring the 1st analysis result information, when additional information is inputted. The step which compounds the aforementioned 1st analysis result information and the 2nd analysis result information, and generates the 3rd analysis result information.
[Claim 2] The information-analysis method according to claim 1 characterized by including the step saved as analysis result information that the information which specifies the information and accumulation frequency which specify the time which analyzed with the aforementioned 2nd analysis result information is used at the time of the next information addition.
[Claim 3] The information-analysis method characterized by providing the following. The step which is the information-analysis method using the correlation rule discovery technique, analyzes the aforementioned additional information and searches for the 2nd analysis result information while verifying the existing analysis result information for additional information and searching for the 1st analysis result information, when information is added and deleted. The step which compounds the analysis result information and the aforementioned 2nd analysis result information which reduce the analysis result information which should be deleted from the aforementioned 1st analysis result information, and are acquired, and generates the 3rd analysis result information.
[Claim 4] Information-analysis equipment characterized by providing the following. A means to be information-analysis equipment using the correlation rule discovery technique, and to input additional information. A means to verify the existing analysis result information for the aforementioned additional information, and to generate the 1st analysis result information when the aforementioned additional information is inputted. A means to analyze the aforementioned additional information and to generate the 2nd analysis result information. A means to compound the aforementioned 1st analysis result information and the aforementioned 2nd analysis result information, and to generate the 3rd analysis result information.
[Claim 5] Information-analysis equipment according to claim 4 characterized by including a means to save as analysis result information that the information which specifies the information and accumulation frequency which specify the time which analyzed with the aforementioned 2nd analysis result information is used at the time of the next information addition.
[Claim 6] Information-analysis equipment characterized by providing the following. A means to be information-analysis equipment using the correlation rule discovery technique, to verify the existing analysis result information for additional information, and to acquire the 1st analysis result information when information is added and deleted. A means to analyze the aforementioned additional information and to acquire the 2nd analysis result information. A means to compound the analysis result information and the aforementioned 2nd analysis result information which reduce the analysis result information which should be deleted from the aforementioned 1st analysis result information, and are acquired, and to generate the 3rd analysis result information.

[Translation done.]

DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[The technical field to which invention belongs] this invention relates to the information-analysis method and equipment which used the correlation rule discovery technique.
[0002]
[Description of the Prior Art] Data mining attracts attention as technology of extracting knowledge from large-scale data **-SU. As the technique of data mining, various technique, such as a decision tree, a neural network, correlation rule discovery, and clustering, is proposed. The feature hidden in data **-SU by such technique is extracted, and the application to various fields, such as marketing, is expected.
[0003] By the basic system, a snapshot is taken periodically and data **-SU generally made into the object of mining uses not the thing under employment but the thing built as another data **-SU (data warehouse). Therefore, usually renewal of data **-SU is performed by adding collectively the data which were not reflected in real time but were added after a fixed period. For this reason, whenever addition of periodical data is performed for grasping the inclination covering whole data **-SU, it is necessary to perform mining about whole data **-SU. Data **-SU set as the object of mining is huge in many cases, and has taken the great execution time to perform mining about whole data **-SU at each time which is addition of data.
[0004] Correlation rule discovery is one of the typical mining technique, and is used as the technique of performing basket analysis in retail trade. Basket analysis is technique as which a customer analyzes the group of the item simultaneously purchased by one transaction, for example, the correlation rule "the customer who buys beer also buys a disposable diaper simultaneously" can be discovered. This processing is performed by the following procedures.
[0005] 1: Ask for the frequency of occurrence according to an item about all transactions.
2: The frequency of occurrence removes the item below the minimum support value.
3: Carry out the self join (SELF JOIN) of this table, and ask for the simultaneous frequency of occurrence of two items.
4: The frequency of occurrence removes the item below the minimum support value.
5: Generate the correlation rule beyond the minimum confidence value about the pair of the extracted item.
[0006] Furthermore, this is repeated and a correlation rule is similarly generated about the group of three or more items. In addition, a user does initial setting of the minimum support value and the minimum confidence value, and a support value and a confidence value are defined as follows about the correlation rule of the form {A1.A2 --An} ->B.
[0007] Support value = (A1.A2 -- number of times of an appearance of An and B) all /number confidence value of transactions =(aluminum.A2 -- number of times of appearance of An and B)/(number of times of an appearance of A1.A2 --An)
The ****** rule between items with the high frequency of occurrence is extracted using these two.
[0008]
[Problem(s) to be Solved by the Invention] It is necessary to search whole data **-SU in the former in quest of the frequency of occurrence according to item, and the frequency of occurrence of the group of an item. Or when the index is created for every item, it is necessary to search the whole index. Moreover, when a large number [the item beyond the minimum support value], the processing which self join operation takes becomes huge. Thus, by correlation rule discovery, the great processing time is taken to analyze to large-scale whole data **-SU.
[0009] That is, whenever the content of data **-SU was added, mining needed to be again performed over whole data **-SU, and the conventional method had taken the great processing time each time.
[0010] Therefore, the purpose of this invention is by using the information-analysis (mining) result performed before the information analysis (mining) only about the portion to which data **-SU was added, and the informational addition to offer the information-analysis method and equipment which extract efficiently the feature included in the content of the newest data **-SU.
[0011]
[Means for Solving the Problem] When this invention is the information-analysis method which used the correlation rule discovery technique and additional information is inputted, The step which analyzes the aforementioned additional information and acquires the 2nd analysis result information while verifying the existing analysis result information for the aforementioned additional information and acquiring the 1st analysis result information, The aforementioned 1st analysis result information and the 2nd analysis result information are compounded, and the information-analysis method characterized by having with the step which generates the 3rd analysis result information is offered.
[0012] When it is the information-analysis method by which the correlation rule discovery technique was used for this invention and information is added and deleted, The step which analyzes the aforementioned additional information and searches for the 2nd analysis result information while verifying the existing analysis result information for additional information and searching for the 1st analysis result information, The analysis result information and the aforementioned 2nd analysis result information which reduce the analysis result information which should be deleted from the aforementioned 1st analysis result information, and are acquired are compounded, and the information-analysis method characterized by generating the 3rd analysis result information is offered.
[0013] Especially, when information is added in correlation rule discovery, this invention carries out mining only of the

additional information, and generates additional information mining information. Verify a correlation rule using the aforementioned additional information to the mining information on the past obtained from mining of the information before an information addition, and the mining information on additional information is compounded to the mining information on past according to this verification result. The incremental information mining method characterized by generating the mining result of whole data **-SU including additional information is offered.
[0014] A means for this invention to be information-analysis equipment which used the correlation rule discovery technique, and to input additional information, A means to verify the existing analysis result information for the aforementioned additional information, and to generate the 1st analysis result information when the aforementioned additional information is inputted, A means to analyze the aforementioned additional information and to generate the 2nd analysis result information, and the aforementioned 1st analysis result information and the aforementioned 2nd analysis result information are compounded, and the information-analysis equipment characterized by providing a means to generate the 3rd analysis result information is offered.
[0015] A means for this invention to verify the existing analysis result information for additional information when it is information-analysis equipment which used the correlation rule discovery technique and information is added and deleted, and to acquire the 1st analysis result information, A means to analyze the aforementioned additional information and to acquire the 2nd analysis result information, the analysis result information which reduces the analysis result information which should be deleted from the aforementioned 1st analysis result information, and is acquired, and the aforementioned 2nd analysis result information are compounded. The information-analysis equipment characterized by providing a means to generate the 3rd analysis result information is offered.
[0016] A means by which this invention adds information in correlation rule discovery, and a new mining means to carry out mining only of the additional information, to extract, and to generate the 1st mining result information, A verification means to verify the past mining result information acquired by mining of the information before adding using the aforementioned additional information, and to generate the 2nd mining result information, The mining result information on the above 2nd and the mining result information on the above 1st which are acquired by this verification means are compounded. The incremental information mining equipment characterized by consisting of synthetic meanses to generate the mining result of whole data **-SU including the aforementioned additional information is offered.
[0017] According to this invention, the feature included in the content of the newest data **-SU is efficiently extracted by using the mining result which performed mining only about additional information and was performed before the informational addition. Therefore, when information is added, it is not necessary to deal with whole large-scale data **-SU, and it becomes possible to accelerate sharply the information mining operation performed daily.
[0018]
[Embodiments of the Invention] Drawing 1 shows the structure of a system which realizes the incremental data-mining method of this invention. According to this, the past mining system and the new mining system are shown. A past mining system contains the original database 11 and the past mining section 12. The original database 11 stores the item data of a large number collected in the past, and the past mining section 12 performs mining to the past data, and it generates the past mining result 13.
[0019] A new mining system is constituted by the additional data generating section 21, the new mining section 22, the verification section 23, and the synthetic section 24. The output of the additional data generating section 21 is connected to the new mining section 22 and the verification section 23, and the output of the new mining section 22 and the verification section 23 is connected to the synthetic section 24.
[0020] Although the new mining section 22 performs the same processing as the conventional mining, it performs mining only about the additional data instead of whole data **-SU. Therefore, mining processing can accelerate sharply compared with the former. The verification section 23 verifies whether the past mining result is succeedingly materialized also to present data **-SU. Specifically, this verification section 23 verifies whether it is realized to additional data as a result of [past] mining (i.e., the past correlation rule). The synthetic section 24 generates information required for judgment of the verification section in next mining while compounding and outputting the result of the new mining section 22 and the verification section 23.
[0021] It is easier to verify whether generally mining of the strange data is carried out, and the knowledge extracted in the past rather than it extracted knowledge is applied to present. For example, in correlation rule discovery, if the group of an item is assumed as knowledge extracted in the past and these will count the frequency which exists in additional data, it is easily verifiable whether the past mining result is applied to additional data. For this reason, it becomes accelerable [mining to whole data **-SU containing the added data].
[0022] (1st operation gestalt) The incremental data-mining method of the 1st operation gestalt of this invention is explained. First, the past mining system which performs data mining about four transactions is explained, referring to the flow chart of drawing 2 . In this example, each transaction is equivalent to one purchase of a consumer, and a unique identification number (TID) is given. In this case, a transaction carries out to four of 100,200,300,400. A, B, C, D, and E express each item. The list of items purchased for every transaction is assumed to be what is shown in Table 1.
[0023] Table 1 TID Item list 100 (A, C, D)
200 (B, C, E)
300 (A, B, C, E)
400 (B, E)
If the above-mentioned item list is read from the original database 11 (S11) and is sent to the past mining section 12, the frequency of occurrence for every item will be called for after this (S12). The frequency of occurrence obtained at this time is shown in Table 2.
[0024] Table 2 item The frequency of occurrence A 2B 3C 3D 1E 3 -- here, the minimum support value is set to 0.3 and the low item of frequency is removed (S13) That is, since the number of transactions is 4, the frequency of occurrence removes less than 1.2 thing. Here, Item D is removed. Self joint is performed about four items which remained (S14), and the group of an item is generated. then, the original transaction data -- the frequency of occurrence of an item group -- asking (S15) -- the frequency of occurrence of an item group becomes as it is shown in Table 3
[0025] A table 3 item group The frequency of occurrence (A, B) 1 (A, C) 2 (A, E) 1 (B, C) 2 (B, E) 3 (C, E) 2 -- in this, since

the frequency of occurrence is under the minimum support value (1.2), (A, B), and (A, E) are removed (S16) Since two or more item groups are obtained, after removal continues processing (S17). That is, processing returns to Step S14 and the self join of a diad is taken (S14). Thereby, three groups of an item are generated. if it asks for the frequency of occurrence from transaction data, the frequency of occurrence of an item group (B, C, E) is 2 -- understanding -- other than this -- being alike -- it turns out that there is no solution A loop is ended here (S17).

[0026] What is necessary is for a confidence value just to decompose the element of the group of an item into the left part and the right-hand side of a rule, in order to generate a correlation rule using the item group detected by the processing so far.

[0027] Confidence value = (number of times of an appearance of left part and the right-hand side) since / (number of times of an appearance of left part) defines, it is set to confidence value =1/3 of confidence value =1-/2B->A of A->B if it attaches, for example (A, B). It becomes the correlation rule by which the thing beyond the minimum confidence value is generated from these. That is, the thing beyond the minimum confidence value is outputted as a mining result (S18). In addition, the portion which serves as a bottleneck on processing in this algorithm is a portion which asks for the item group beyond the minimum support value, and targets even the place which outputs the item group below the minimum support value as a mining result. Therefore, as shown in Table 4, let the mining result about this example be each frequency of occurrence with an item group.

[0028] Table 4 item group Frequency of occurrence (A, C) 2 (B, C) 2 (B, E) 3 (C, E) 2 (B, C, E) Operation of the new mining section is explained about the case where there are 2, next additional data, referring to the flow chart of drawing 3 . The additional data to above-mentioned data **-SU shall be shown in Table 5.

[0029] Table 5 TID Item list 500 (A, B, C)
600 (A, C, E)
700 (B, E, F)
800 (A, B, F)
An input of this additional data asks for the frequency of occurrence about this additional data (S22). (S21) The frequency of occurrence obtained at this time is shown in Table 6.

[0030] Table 6 item The frequency of occurrence A 3B 3C 2E 2F 2 -- here, the minimum support value is set to 0.3 and the low item of frequency is removed (S23) That is, since the number of transactions is 4, the frequency of occurrence removes less than 1.2 thing. Here, since there is no item for removal, self joint is performed about five items (S24), and an item group is generated. then, the original transaction data -- the frequency of occurrence of an item group -- asking (S25) -- the frequency of occurrence of an item group becomes as it is shown in Table 7

[0031] Table 7 item The frequency of occurrence (A, B) 2 (A, C) 2 (B, F) 2 (E, F) 1 -- in this, since the frequency of occurrence is under the minimum support value, (E, F) are removed (S26) Thereby, three item groups are generated. It turns out that it turns out that the frequency of occurrence of these item group is 2 when it asks for the frequency of occurrence from transaction data, and there is no solution in addition to it. A loop is ended here (S17). And the group of the item beyond the minimum support value is chosen (S28). Thereby, the item group shown in Table 8 and its frequency of occurrence are obtained. This is equivalent to the result related only with additional data.

[0032] Table 8 item Frequency of occurrence (A, B) 2 (A, C) 2 (B, F) Mining of 2, next whole data **-SU which added additional data is explained. First, it explains that the right mining result is not obtained only by totaling the mining result before an addition, and the mining result about additional data simply.

[0033] If the mining result of additional data shown in the mining result and Table 8 before the addition shown in Table 4 is totaled, since the number of transactions will be set to 8, if it is the minimum support value 0.3, two item groups which frequency shows in Table 9 as 2.4 or more item groups will be obtained.

[0034] Table 9 item The frequency of occurrence (A, C) 4 (B, E) 3 -- on the other hand -- additional data -- beforehand -- original data **-SU -- in addition, if mining is performed from the whole, the result which frequency shows in Table 10 as a group of 2.4 or more items will be obtained

[0035] A Table 10 item The frequency of occurrence (A, B) 3 (A, C) 4 (B, C) 3 (B, E) 4 (C, E) It turns out that totaling the result which divided and carried out mining in five results obtained by carrying out mining on the whole only by totaling the mining result addition before and after an addition, and being obtained so that it may understand, if three tables 9 are compared with Table 10 is set only to two, and three information is lost.

[0036] The method of this invention verifies the mining result before an addition to additional data, and compounds the mining result of additional data to this. This technique is explained with reference to the flow chart of drawing 4 and drawing 5 below.

[0037] It asks as a result of [over the data before an addition (TID=100-400)] mining (i.e., the past mining result) (S31). This mining result is the same as Table 4. About these, it verifies to additional data (TID=500-800). That is, the frequency of occurrence in additional data is computed (S32), and it is added to the frequency to which an item group appears in additional data (S33). The mining result which added the verification result comes to be shown in Table 11.

[0038] A Table 11 item The frequency of occurrence (A, C) 2+2=4 (B, C) 2+1=3 (B, E) 3+1=4 (C, E) 2+1=3 (B, C, E) 2+0=2 (A, C), (B, C), (B, E), and (C, E) are compared with the minimum support value (S34). Since these item group is beyond the minimum support value, these are passed to the synthetic section 24 (S35).

[0039] Moreover, the mining result only of additional data is as having been shown in Table 8, and as shown in the following table 12, three item groups are obtained. This is passed to the synthetic section 24.

[0040] Table 12 item Frequency of occurrence (A, B) 2 (A, C) 2 (B, F) In 2 composition sections 24, as shown in the flow chart of drawing 5 , the result (S41) of the new mining section 22 and the data (S42) of the verification section 23 are compounded, and an additional mining result is generated. In this composition, it is judged whether the rule generated exists in both the continuation from the mining result of past and a new mining result (S43). If this judgment is NO, it will be judged whether it exists only in the output of the new mining section (S44). If a rule exists in both, it will be outputted as continuation (S45). If a rule exists only in the new mining section, it will be outputted as a new output (S46). Continuation / new distinction is written together by each rule at this time. The result of composition becomes as it is shown in Table 13.

[0041] A table 13 item group The frequency of occurrence (A, C) 4 Continuation (B, C) 3 Continuation (B, E) 4 Continuation (C, E) 3 Continuation (A, B) 2 New (B, F) 2 new -- all of five rules found when the mining result of this addition was compared with the result (Table 10) which performed mining by the whole which added additional data and mining was performed on the whole are contained, and they are still (B, F) more newly extracted by the technique of this invention The

capacity for this to extract the feature continuously generated in the technique of this invention is equivalent to the result which performed mining at whole data **-SU, and it is shown that there is capacity to extract the feature (B, F) which is included only about new data in addition to it.
[0042] Although the case where data were added only once was explained above, data are added continuously and the case where mining is performed to whenever [the] is explained. The structure of a system in this case is shown in drawing 6 . According to this, the initial mining system and the new mining system are shown. An initial mining system contains the initial database 31 and the initial mining section 32. The initial database 31 stores the item data of a large number collected in early stages, and the initial mining section 32 performs mining to early data, and it generates the early mining result 33.
[0043] A new mining system is constituted by the additional data generating section 21, the new mining section 22, the verification section 23, and the synthetic section 24 like drawing 1 . According to this system, the output of the synthetic section 24 is used for next time as a mining result.
[0044] For example, when data are added once every month and mining is performed to additional data per moon, it is thought that remarkable dispersion exists in a monthly mining result. On the other hand, if mining is performed to whole data **-SU after adding data, only the rule that frequency is high will be extracted through the whole.
[0045] In the former, in order to have extracted the rule of these both, two mining, mining and the whole mining, about additional data needed to be performed. By the technique of this invention, it becomes possible to search for the rule that frequency is high, efficiently through the whole, without performing mining to the whole on the basis of mining to additional data.
[0046] Then, the example in which data are continuously added to below is explained. Time which performs the first mining is set to 0, and suppose that there was addition of data at time 1, 2, 3, and 4, respectively. The data number of cases in time 0 and the number of cases of the data added in each time may be 1000 affairs, respectively. The minimum support value shall extract the rule of 100 or more frequency in 0.1, i.e., the data added in each time.
[0047] As a result of performing mining of additional data about time 0-4, it is assumed that the frequency within the data added in each time about six sorts of rules as shown in Table 14 was obtained.
[0048]
table 14 Time 0 1 2 3 4 A rule 1 200 160 180 150 140 A rule 2 150 40 30 10 10 A rule 3 120 120 80 90 120 A rule 4 100 60 110 70 100 Rule 580 130 120 140 150 Rule 6 40 50 150 120 If mining is performed about 90, i.e., the data added at each time, 100 or more rules will be acquired for frequency as a result. That is, an underline portion is outputted as a mining result in Table 14.
[0049] Next, after adding data in each time, the case where mining is performed about the whole is explained. The frequency of each rule serves as an accumulation value of the frequency by the time, and becomes as it is shown in Table 15.
[0050]
table 15 Time 0 1 2 3 4 A rule 1 200 360 540 690 830 A rule 2 150 190 220 230 240 A rule 3 120 240 320 410 530 A rule 4 100 160 270 340 440 Rule 5 80 210 330 470 620 Rule 6 40 90240 360 450 -- this case -- time 0 -- 500 or more rules are outputted at 200 or more and time 2 in 100 or more and time 1, and are outputted as a mining result at 400 or more and time 4 in 300 or more and time 3 That is, an underline portion is outputted as a result in Table 15.
[0051] The technique of this invention is set in the synthetic section, as shown in drawing 7 , and as a mining result of each time, the following procedures shall generate three information, a rule, a start time, and accumulation frequency, and it shall save and reuse it.
[0052] First, it is judged whether the rule is included in the accumulation mining result 33 (S51). If this judgment is YES (i.e., if it is the rule included in the past mining result), the frequency of the additional data of the present time will be applied to the accumulation frequency of the past mining result, a rule will be outputted (S54), and a start time will presuppose that it remains as it is (S55).
[0053] If a judgment at Step 51 is NO, namely, if it is the rule which is not included in the past mining result and the frequency of the additional data of the present time is higher than the minimum support value, a rule will be outputted for accumulation frequency as frequency of the additional data of the present time (S52), and let a start time be the present time (S53).
[0054] If this technique is applied to the above-mentioned example, the output of mining in each time will become as it is shown in the following table 16.
[0055]
table 16 A rule A start time Accumulation frequency time 0 Rule 1 0 200 A rule 2 0 150 A rule 3 0 120 A rule 4 0 100 Time 1 Rule 1 0 200+60=360 Rule 2 0 150+40=190 Rule 3 0 120+120=240 Rule 40 100+60=160 Rule 5 1130 time 2 Rule 10 180= 360+540 rule 2 0190+30=220 Rule 30 240+80=320 A rule 40 160+110=270 A rule 5 1130+120=250 Rule 6 2 I50 time 3 rule 1 0540+150=690 A rule 2 0 220+10=230 A rule 3 0 320+90=410 Rule 4 0 270+70=340 Rule 5 1 250+140=390 A rule 6 2 150+120=270 Time 4 Rule 1 0 690+140=830 Rule 2 0 230+10=240 Rule 3 0 410+120=530 Rule 4 0 100= 340+440 rule 5 1 390+150=540 Rule 6 2270+90=360 -- when it does in this way, the rule which has the frequency beyond the minimum support value also at once in the data added in a certain time will be outputted as a mining result all the time after that That is, all the results obtained by carrying out mining about whole data **-SU in arbitrary time are included in this list.
[0056] In addition, by this technique, since it increases whenever a mining result adds data, the execution time of mining may increase. The method of removing the rule outputted as the improvement when the ratio of accumulation frequency becomes below fixed is also considered. For example, if the ratio of accumulation frequency becomes 0.05 or less, supposing it will remove a rule from a result, a rule 2 will be removed at time 4. Such judgment is easily calculable if the number of transactions added at a start time and each time is held.
[0057] (2nd operation gestalt) Although the 1st operation gestalt describes the case where data **-SU is added, usage which sets constant the period of the data stored in data **-SU may be carried out like the past one year. In this case, it is necessary to remove the data which separated from the period whenever it added new data, and to take removal into consideration also about maintenance of a mining result.
[0058] Below, the periodic increment mining system according to the 2nd operation gestalt of this invention is explained with reference to drawing 8 .
[0059] According to the composition of drawing 8 , the mining result 41 classified by time is added to the system of drawing 6

. The same data as the example used with the 1st operation gestalt explain this system. That is, the same thing as Table 14 is used for the frequency of occurrence of the rule 1-6 in time 0-5.
[0060] Here, a period shall hold 3, i.e., the past 3 times of data. The mining result of whole data **-SU when setting a period to 3 is shown in Table 17.
[0061]
table 17 Time 0 1 2 3 4 A rule 1 200 360 540 490 470 A rule 2 150 190 220 80 50 A rule 3 120 240 320 290 290 A rule 4 100 160 270 240 280 rule 580 210 330 390 410 Rule 6 40 90240 Frequency is outputted at 320 360, in this case time 0, and, henceforth [200 or more and time 3], 300 or more rules are outputted as a mining result at 100 or more and time 1. That is, an underline portion is outputted as a result in the above-mentioned table 17.
[0062] Below, in a period 3, the technique of searching for the whole mining result from the mining result of an additional portion and the past mining result is explained with reference to the flow chart of drawing 9 .
[0063] Till time 2, it is the same as that of the 1st operation gestalt, the data of time 0 are removed at the time of time 3, the data of time 3 are added, at time 4, the data of time 1 are deleted and the data of time 4 are added. In addition to holding the information which specifies the content of a rule, a start time, and accumulation frequency about the rule realized about whole data **-SU like the 1st operation gestalt as a mining result, it shall hold as a result of [about the additional data in each time / 41] mining (i.e., the frequency of occurrence in the additional data of the rule which it is at the data addition-time and is outputted). The procedure in each time is performed as shown in the flow chart of drawing 9 .
[0064] First, it is judged whether the rule is included in the accumulation mining result 33 (S61). If this judgment is YES (i.e., if a rule is a rule included in the past mining result), it will be judged for a start time whether it is before before 1 period (S62). If this judgment is YES, accumulation frequency will be computed by the frequency of the frequency + present time at the time of the last accumulation frequency-deletion (S63). That is, the accumulation mining result of a fixed period reduces the mining result of the period which should delete an accumulation mining result from the mining result verified and obtained by additional data, and is searched for by compounding an additional mining result. A start time is made into 1 period front +1 (S64).
[0065] If a judgment at Step S61 is YES and a judgment at Step S62 is NO, accumulation frequency will be called for with the frequency of the last accumulation frequency + present time (S65), and let a start time be a value as it is (S66).
[0066] In the rule which is not included in the past mining result if the judgment of Step S61 is NO, if the frequency in the additional data of existing time is higher than the minimum support value, a rule will be outputted as frequency [in / the additional data of the present time / for accumulation frequency] (S67), and let a start time be the present time (S68).
[0067] The mining result in each time at the time of considering as the period 3 according to the above-mentioned procedure is shown in Table 18.
[0068]
table 18 A rule Start time Accumulation frequency time 0 Rule 1 0 200 A rule 2 0 150 A rule 3 0 120 A rule 4 0 100 Time 1 Rule 1 0 200+60=360 Rule 20 40= 150+190 rule 3 0 120+120=240 A rule 4 0 100+60=160 Rule 5 1130 time 2 Rule 10 180= 360+540 rule 2 0190+30=220 Rule 30 240+80=320 Rule 4 0 160+110=270 Rule 5 1 130+120=250 Rule 6 2 150 Time 3 Rule 1 1540+150-200=490 A rule 2 1 220+10-150=80 Rule 3 1 320+90-120=290 Rule 4 1 270+70-100=240 Rule 5 1 250+140=390 Rule 6 2 150+120=270 Time 4 Rule 1 2 490+140-160=470 A rule 2 2 80+10-40=50 Rule 32 290+120-120=290 Rule 4 2 240+100-60=280 Rule 5 2 390+150-130=410 Rule 6 2 The mining result outputted [in / this method / for whether it being 90= 270+360 Ming et al.] includes the mining result which followed whole data **-SU. Moreover, it is also easy to delete the rule from which frequency became below fixed from a mining result like the 1st operation gestalt.
[0069] The whole mining is performed by compounding the mining result which verified and obtained the past mining result about additional data, and the mining result about additional data, without accessing past data **-SU according to this invention, as mentioned above, when there are an addition and deletion of data.
[0070]
[Effect of the Invention] It is effective in order for mining of the whole database to become possible and to perform mining of large-scale data efficiently by compounding the mining result of data **-SU before adding with mining of the data added, without according to this invention carrying out mining of whole data **-SU when data are added to data **-SU.
[0071] Moreover, it is effective in order for mining of whole data **-SU to become possible similarly using the past mining result in a periodic database which deletes the data of the oldest time at the time of addition of data and to perform mining of large-scale data efficiently.

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]
[Drawing 1] The block diagram of an incremental mining system according to 1 operation gestalt of this invention.
[Drawing 2] The flow chart explaining the incremental mining method for obtaining the mining result of the past of this invention.
[Drawing 3] The flow chart explaining the incremental mining method for obtaining the new mining result according to the 1st operation gestalt.
[Drawing 4] The flow chart explaining the verification section used by new mining of the 1st operation gestalt.
[Drawing 5] The flow chart explaining the synthetic section used by new mining of the 1st operation gestalt.
[Drawing 6] The block diagram of the incremental mining system using an initial mining result.
[Drawing 7] The flow chart explaining the synthetic section in the mining system of drawing 6 .
[Drawing 8] The block diagram of an incremental mining system according to the 2nd operation gestalt of this invention.
[Drawing 9] The flow chart explaining the synthetic section in the mining system of drawing 8 .
[Description of Notations]
11 -- Field database
12 -- Past mining section
13 -- The past mining result
21 -- Additional data section
22 -- New mining section
23 -- Verification section
24 -- The synthetic section
31 -- Initial database
32 -- Initial mining section

[Translation done.]

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（A）**　(11)特許出願公開番号
特開2001－344259
（P2001－344259A）
(43)公開日　平成13年12月14日(2001.12.14)

(51)Int.Cl.'		識別記号	FI		テーマコード'(参考)
G06F	17/30	220	G06F	17/30	220Z　5B075
	12/00	510		12/00	510A　5B082
		512			512

審査請求　有　　請求項の数6　OL　（全 12 頁）

(21)出願番号　特願2000－162080(P2000－162080)

(22)出願日　平成12年5月31日(2000.5.31)

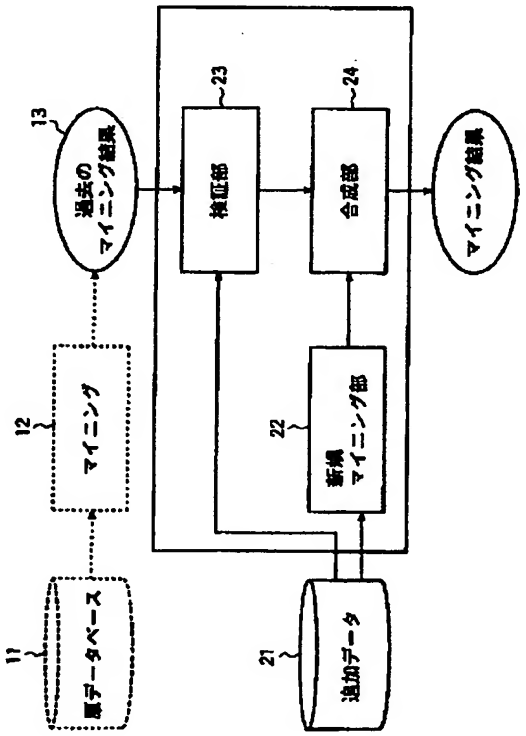
（出願人による申告）国等の委託研究の成果に係る特許出願（平成12年度通産省委託事業「エネルギー使用合理化電子計算機技術開発」委託研究、産業活力再生特別措置法第30条の適用を受けるもの）

(71)出願人　000003078
株式会社東芝
東京都港区芝浦一丁目１番１号
(72)発明者　小柳　滋
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝研究開発センター内
(72)発明者　酒井　浩
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝研究開発センター内
(74)代理人　100058479
弁理士　鈴江　武彦　（外６名）

最終頁に続く

(54)【発明の名称】　情報分析方法および装置

(57)【要約】
【課題】本発明は、データの追加・削除があったとき、データマイニングを高速化するインクリメンタルマイニング方法を提供する。
【解決手段】データマイニングの手法である相関規則発見において、過去のマイニング結果を保存しておき、データの追加・削除があったときに過去のデータベースにアクセスすることなく、過去のマイニング結果を追加データについて検証したものと、追加データに関するマイニング結果を合成することにより全体のマイニングを行う。



【特許請求の範囲】
【請求項１】　相関規則発見手法を用いた情報分析方法であって、
追加情報が入力された際、既存の分析結果情報を前記追加情報にて検証して第１分析結果情報を得るとともに前記追加情報を分析して第２分析結果情報を得るステップと、前記第１分析結果情報と第２分析結果情報とを合成し、第３分析結果情報を生成するステップとを有することを特徴とする情報分析方法。
【請求項２】　前記第２分析結果情報とともに、分析を行った時刻を特定する情報および累積頻度を特定する情報を次の情報追加時に利用する分析結果情報として保存するステップを含むことを特徴とする請求項１記載の情報分析方法。
【請求項３】　相関規則発見手法を用いた情報分析方法であって情報が追加および削除された際、既存の分析結果情報を追加情報にて検証して第１分析結果情報を求めるとともに前記追加情報を分析して第２分析結果情報を求めるステップと、前記第１分析結果情報から削除すべき分析結果情報を減じて得られる分析結果情報と前記第２分析結果情報を合成して、第３分析結果情報を生成するステップを有することを特徴とする情報分析方法。
【請求項４】　相関規則発見手法を用いた情報分析装置であって、
追加情報を入力する手段と、前記追加情報が入力された際、既存の分析結果情報を前記追加情報にて検証して第１分析結果情報を生成する手段と、前記追加情報を分析して第２分析結果情報を生成する手段と、前記第１分析結果情報と前記第２分析結果情報とを合成し、第３分析結果情報を生成する手段とを具備することを特徴とする情報分析装置。
【請求項５】　前記第２分析結果情報とともに、分析を行った時刻を特定する情報および累積頻度を特定する情報を次の情報追加時に利用する分析結果情報として保存する手段を含むことを特徴とする請求項４記載の情報分析装置。
【請求項６】　相関規則発見手法を用いた情報分析装置であって情報が追加および削除された際、既存の分析結果情報を追加情報にて検証して第１分析結果情報を得る手段と、前記追加情報を分析して第２分析結果情報を得る手段と、前記第１分析結果情報から削除すべき分析結果情報を減じて得られる分析結果情報と前記第２分析結果情報を合成して、第３分析結果情報を生成する手段とを具備することを特徴とする情報分析装置。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、相関規則発見手法を用いた情報分析方法および装置に関する。
【０００２】
【従来の技術】大規模データベースから知識を抽出する技術としてデータマイニングが注目されている。データマイニングの手法としては、決定木、ニューラルネット、相関規則発見、クラスタリングなど様々な手法が提案されている。これらの手法によりデータベースの中に隠されている特徴を抽出し、マーケテイングなどのさまざまな分野への応用が期待されている。
【０００３】一般にマイニングの対象とするデータベースは基幹システムで運用中のものではなく、定期的にスナップショットをとり、別のデータベース(データウエアハウス)として構築したものを利用する。そのため、データベースの更新はリアルタイムに反映されず、一定期間の後に追加されたデータを一括して追加することにより行われるのが通常である。このため、データベース全体にわたる傾向を把握するには定期的なデータの追加が行われる度にデータベース全体についてマイニングを行う必要がある。マイニングの対象となるデータベースは膨大な場合が多く、データの追加の都度にデータベース全体についてマイニングを実行するのには多大な実行時間を要している。
【０００４】相関規則発見は代表的なマイニング手法の一つであり、小売業におけるバスケット分析を行う手法として利用されている。バスケット分析とは、顧客が１トランザクションで同時に購入するアイテムの組を分析する手法であり、例えば「ビールを買う顧客は同時に紙おむつも買う」というような相関規則が発見できる。この処理は以下のような手順で行われる。
【０００５】１：全トランザクションについてアイテム別に出現頻度を求める。
２：出現頻度が最小サポート値以下のアイテムを除去する。
３：この表をセルフジョイン（SELF JOIN）して２つのアイテムの同時出現頻度を求める。
４：出現頻度が最小サポート値以下のアイテムを除去する。
５：抽出されたアイテムの対について最小コンフィデンス値以上の相関規則を生成する。
【０００６】さらに、これを繰り返して、３個以上のアイテムの組についても同様に相関規則を生成する。なお、最小サポート値、最小コンフィデンス値はユーザが初期設定するものであり、｛Ａ１．Ａ２…Ａｎ｝→Ｂという形式の相関規則に関して、サポート値、コンフィデンス値は以下のように定義される。
【０００７】サポート値＝（Ａ１．Ａ２…Ａｎ、Ｂの出現回数）／全トランザクション数
コンフィデンス値＝（Ａ１．Ａ２…Ａｎ、Ｂの出現回数）／（Ａ１．Ａ２…Ａｎの出現回数）
この２つを用いて、出現頻度の高いアイテム間の相関測則が抽出される。
【０００８】
【発明が解決しようとする課題】従来では、アイテム別

の出現頻度、およびアイテムの組の出現頻度を求めるにはデータベース全体をサーチする必要がある。あるいは、アイテム毎にインデックスが作成されている場合には、インデックス全体をサーチする必要がある。また、最小サポート値以上のアイテムが多数ある場合には、セルフジョイン操作に要する処理が膨大となる。このように相関規則発見では大規模なデータベース全体に対して分析するのに、多大な処理時間を要する。

【0009】即ち、従来の方法では、データベースの内容が追加される度に、データベース全体にわたって再度マイニングを行う必要があり、その都度多大な処理時間を要していた。

【0010】従って、本発明の目的は、データベースの追加された部分のみに関する情報分析（マイニング）、および情報の追加以前に行われた情報分析（マイニング）結果を利用することにより、最新のデータベースの内容に含まれる特徴を効率よく抽出する情報分析方法および装置を提供することにある。

【0011】

【課題を解決するための手段】本発明は、相関規則発見手法を用いた情報分析方法であって、追加情報が入力された際、既存の分析結果情報を前記追加情報にて検証して第1分析結果情報を得るとともに前記追加情報を分析して第2分析結果情報を得るステップと、前記第1分析結果情報と第2分析結果情報とを合成し、第3分析結果情報を生成するステップと有することを特徴とする情報分析方法を提供する。

【0012】本発明は、相関規則発見手法を用いた情報分析方法であって情報が追加および削除された際、既存の分析結果情報を追加情報にて検証して第1分析結果情報を求めるとともに前記追加情報を分析して第2分析結果情報を求めるステップと、前記第1分析結果情報から削除すべき分析結果情報を減じて得られる分析結果情報と前記第2分析結果情報を合成して、第3分析結果情報を生成することを特徴とする情報分析方法を提供する。

【0013】特に本発明は、相関規則発見において情報が追加されたとき、追加情報のみをマイニングして追加情報マイニング情報を生成し、情報追加以前の情報のマイニングより得られた過去のマイニング情報に対して前記追加情報により相関規則の検証を行い、この検証結果に従って過去のマイニング情報に追加情報のマイニング情報を合成して、追加情報を含むデータベース全体のマイニング結果を生成することを特徴とするインクリメンタル情報マイニング方法を提供する。

【0014】本発明は、相関規則発見手法を用いた情報分析装置であって、追加情報を入力する手段と、前記追加情報が入力された際、既存の分析結果情報を前記追加情報にて検証して第1分析結果情報を生成する手段と、前記追加情報を分析して第2分析結果情報を生成する手段と、前記第1分析結果情報と前記第2分析結果情報とを合成し、第3分析結果情報を生成する手段とを具備することを特徴とする情報分析装置を提供する。

【0015】本発明は、相関規則発見手法を用いた情報分析装置であって、情報が追加および削除された際、既存の分析結果情報を追加情報にて検証して第1分析結果情報を得る手段と、前記追加情報を分析して第2分析結果情報を得る手段と、前記第1分析結果情報から削除すべき分析結果情報を減じて得られる分析結果情報と前記第2分析結果情報を合成して、第3分析結果情報を生成する手段とを具備することを特徴とする情報分析装置を提供する。

【0016】本発明は、相関規則発見において情報を追加する手段と、追加情報のみをマイニングして抽出し、第1のマイニング結果情報を生成する新規マイニング手段と、追加される以前の情報のマイニングにより得られた過去マイニング結果情報を前記追加情報により検証して第2のマイニング結果情報を生成する検証手段と、この検証手段により得られる前記第2のマイニング結果情報と前記第1のマイニング結果情報とを合成して、前記追加情報を含むデータベース全体のマイニング結果を生成する合成手段とで構成されることを特徴とするインクリメンタル情報マイニング装置を提供する。

【0017】本発明によると、追加情報のみについてマイニングを行い、情報の追加以前に行われたマイニング結果を利用することにより、最新のデータベースの内容に含まれる特徴が効率よく抽出される。従って、情報が追加されるときに大規模データベース全体を取扱う必要がなく、日常的に行われる情報マイニング操作を大幅に高速化することが可能となる。

【0018】

【発明の実施の形態】図1は、この発明のインクリメンタルデータマイニング方法を実現するシステムの構成を示している。これによると、過去マイニング系と新規マイニング系が示されている。過去マイニング系は、原データベース11と過去マイニング部12とを含む。原データベース11は過去に収集された多数のアイテムデータを格納しており、過去マイニング部12は過去のデータに対してマイニングを行い、過去のマイニング結果13を生成する。

【0019】新規マイニング系は、追加データ発生部21，新規マイニング部22，検証部23および合成部24により構成される。追加データ発生部21の出力は新規マイニング部22および検証部23に接続され、新規マイニング部22および検証部23の出力は合成部24に接続される。

【0020】新規マイニング部22は、従来のマイニングと同様の処理を行うが、データベース全体ではなく、追加データのみについてマイニングを行う。従って、マイニング処理が、従来と比べて大幅に高速化できる。検証部23は過去のマイニング結果が現在のデータベース

に対しても引き続き成立するか否かを検証するものである。具体的には、この検証部２３は、過去のマイニング結果、即ち過去の相関規則が追加データに対して成り立つかを検証する。合成部２４は新規マイニング部２２と検証部２３の結果を合成して出力するとともに、次回のマイニングにおける検証部の判断に必要な情報を生成する。

【００２１】一般に未知のデータをマイニングして知識を抽出するよりも、過去に抽出された知識が現在に当てはまるかどうかを検証する方が容易である。例えば相関規則発見においては、過去に抽出された知識としてアイテムの組を想定すると、これらが追加データ中に存在する頻度を数えれば追加データに対して過去のマイニング結果が当てはまるか否かを容易に検証することができる。このため、追加されたデータを含むデータベース全体に対するマイニングの高速化が可能となる。

【００２２】（第１の実施形態）本発明の第１の実施形態のインクリメンタルデータマイニング方法を説明する。先ず、４つのトランザクションについてデータマイニングを行う過去マイニング系を図２のフローチャートを参照しながら説明する。この例では、各トランザクションは一回の消費者の購入に相当し、ユニークな識別番号(ＴＩＤ)が与えられる。この場合、トランザクションは、１００、２００、３００、４００の4つとする。A、B、C、D、Eは個々のアイテムを表す。各トランザクション毎に購入したアイテムのリストは表１に示すものと仮定する。

【００２３】表1

ＴＩＤ	アイテムリスト
１００	（A、C、D）
２００	（B、C、E）
３００	（A、B、C、E）
４００	（B、E）

上記のアイテムリストが原データベース１１から読み出され（Ｓ１１）、過去マイニング部１２に送られると、これからアイテム毎の出現頻度が求められる（Ｓ１２）。このときに得られる出現頻度が表２に示される。

【００２４】表2

アイテム	出現頻度
A	2
B	3
C	3
D	1
E	3

ここで、最小サポート値を０．３とし、頻度の低いアイテムを除去する（Ｓ１３）。すなわちトランザクション数が４であるので、出現頻度が１．２未満のものを除去する。ここではアイテムDが除去される。残った４つのアイテムに関してセルフジョイントを行い（Ｓ１４）、アイテムの組を生成する。この後、元のトランザクションデータよりアイテム組の出現頻度を求める（Ｓ１５）と、アイテム組の出現頻度は表３のようになる。

【００２５】表3

アイテム組	出現頻度
（A、B）	1
（A、C）	2
（A、E）	1
（B、C）	2
（B、E）	3
（C、E）	2

この中で、（A、B）、（A、E）は出現頻度が最小サポート値（１．２）未満であるので除去する（Ｓ１６）。除去後も、複数個のアイテム組が得られるので処理は継続する（Ｓ１７）。即ち、処理はステップＳ１４に戻り、２つ組のセルフジョインが取られる（Ｓ１４）。これにより、アイテムの３つの組が生成される。トランザクションデータより出現頻度を求めるとアイテム組（B、C、E）の出現頻度が２であることがわかり、それ以外には解がないことが分かる。ここでループは終了する（Ｓ１７）。

【００２６】ここまでの処理により検出されたアイテム組を用いて相関規則を生成するには、アイテムの組の要素をコンフィデンス値により規則の左辺と右辺に分解すればよい。

【００２７】コンフィデンス値＝(左辺と右辺の出現回数)/(左辺の出現回数)により定義されているので、例えば（A、B)については

A→Bのコンフィデンス値＝1/2

B→Aのコンフィデンス値＝1/3

となる。これらより、最小コンフィデンス値以上のものが生成される相関規則となる。即ち、最小コンフィデンス値以上のものがマイニング結果として出力される（Ｓ１８）。なお、本アルゴリズムにおいて処理上のボトルネックとなる部分は最小サポート値以上のアイテム組を求める部分であり、マイニング結果としては最小サポート値以下のアイテム組を出力するところまでを対象とする。従って、この例に関するマイニング結果は表４に示すように、アイテム組と、それぞれの出現頻度とする。

【００２８】表4

アイテム組	出現頻度
（A、C）	2
（B、C）	2
（B、E）	3
（C、E）	2
（B、C、E）	2

次に、追加データがある場合について新規マイニング部の動作を図３のフローチャートを参照しながら説明する。上記のデータベースに対する追加データは表５に示すものとする。

【００２９】表5

TID	アイテムリスト
500	(A、B、C)
600	(A、C、E)
700	(B、E、F)
800	(A、B、F)

この追加データが入力されると（S21）、この追加データについて出現頻度が求められる（S22）。このときに得られる出現頻度が表6に示される。

【0030】表6

アイテム	出現頻度
A	3
B	3
C	2
E	2
F	2

ここで、最小サポート値を0．3とし、頻度の低いアイテムを除去する（S23）。すなわちトランザクション数が4であるので、出現頻度が1．2未満のものを除去する。ここでは除去対象アイテムがないので、5つのアイテムに関してセルフジョイントを行い（S24）、アイテム組を生成する。この後、元のトランザクションデータよりアイテム組の出現頻度を求める（S25）と、アイテム組の出現頻度は表7のようになる。

【0031】表7

アイテム	出現頻度
(A、B)	2
(A、C)	2
(B、F)	2
(E、F)	1

この中で、(E、F)は出現頻度が最小サポート値未満であるので除去する（S26）。これにより、3つのアイテム組が生成される。トランザクションデータより出現頻度を求めるとこれらアイテム組の出現頻度が2であることがわかり、それ以外には解がないことが分かる。ここでループは終了する（S17）。そして最小サポート値以上のアイテムの組が選ばれる（S28）。これにより、表8に示すアイテム組とその出現頻度が得られる。これは追加データのみに関する結果に相当する。

【0032】表8

アイテム	出現頻度
(A、B)	2
(A、C)	2
(B、F)	2

次に、追加データを加えたデータベース全体のマイニングについて説明する。まず、単純に追加前のマイニング結果と追加データに関するマイニング結果を合計するだけでは正しいマイニング結果が得られないことを説明する。

【0033】表4に示した追加前のマイニング結果と表8に示した追加データのマイニング結果を合計すると、トランザクション数は8となるので最小サポート値0．3とすると頻度が2．4以上のアイテム組として表9に示す2つのアイテム組が得られる。

【0034】表9

アイテム	出現頻度
(A、C)	4
(B、E)	3

一方、追加データを予め元のデータベースに加えて、全体からマイニングを行うと、頻度が2．4以上のアイテムの組として表10に示す結果が得られる。

【0035】表10

アイテム	出現頻度
(A、B)	3
(A、C)	4
(B、C)	3
(B、E)	4
(C、E)	3

表9と表10を比べればわかるように、追加前と追加後のマイニング結果を合計するだけでは、全体でマイニングして得られた5つの結果の中で、分割してマイニングした結果を合計して得られるのは2つのみとなり、3つの情報が失われることがわかる。

【0036】本発明の方法は、追加前のマイニング結果を追加データに対して検証し、これに追加データのマイニング結果を合成するというものである。以下この手法について図4および図5のフローチャートを参照して説明する。

【0037】追加前のデータ（TID＝100～400）に対するマイニング結果、即ち過去のマイニング結果が求められる（S31）。このマイニング結果は、表4と同じである。これらについて、追加データ（TID＝500～800）に対して検証を行う。すなわち、追加データ中の出現頻度が算出され（S32）、そしてアイテム組が追加データの中に現れる頻度に加算される（S33）。検証結果を加えたマイニング結果は、表11に示すようになる。

【0038】表11

アイテム	出現頻度
(A、C)	2＋2＝4
(B、C)	2＋1＝3
(B、E)	3＋1＝4
(C、E)	2＋1＝3
(B、C、E)	2＋0＝2

(A、C)、(B、C)、(B、E)、(C、E)は最小サポート値と比較される（S34）。これらアイテム組は最小サポート値以上なので、これらは合成部24に渡される（S35）。

【0039】また、追加データのみに対するマイニング結果は、表8に示した通りであり、下表12に示すように3個のアイテム組が得られる。これが合成部24に渡

される。

【0040】表12

アイテム	出現頻度
(A、B)	2
(A、C)	2
(B、F)	2

合成部24では、図5のフローチャートに示すように新規マイニング部22の結果（S41）と検証部23のデータ（S42）とを合成し、追加のマイニング結果を生成する。この合成において、生成される規則が過去のマイニング結果からの継続と新規マイニング結果の両方に存在するかが判定される（S43）。この判定がNOであれば、新規マイニング部の出力のみに存在するかが判定される（S44）。規則が両方に存在すれば、継続として出力される（S45）。規則が新規マイニング部にのみに存在すれば、新規出力として出力される（S46）。このとき、それぞれの規則に継続/新規の区別が併記される。合成の結果は表13のようになる。

【0041】表13

アイテム組	出現頻度	
(A、C)	4	継続
(B、C)	3	継続
(B、E)	4	継続
(C、E)	3	継続
(A、B)	2	新規
(B、F)	2	新規

この追加のマイニング結果と、追加データを加えた全体でマイニングを行った結果（表10）とを比べてみると、全体でマイニングを行った場合に見つかった5個の規則はすべて含まれており、さらに（B、F）が本発明の手法で新たに抽出されている。これは、本発明の手法において継続的に発生する特徴を抽出する能力はデータベース全体でマイニングを行った結果と等価であり、それに加えて新規データのみについて含まれている特徴（B、F）を摘出する能力があることを示している。

【0042】以上ではデータが1度だけ追加される場合について説明したが、データが継続的に追加され、その度にマイニングを行う場合について説明する。この場合のシステムの構成が図6に示されている。これによると、初期マイニング系と新規マイニング系が示されている。初期マイニング系は、初期データベース31と初期マイニング部32とを含む。初期データベース31は初期に収集された多数のアイテムデータを格納しており、初期マイニング部32は初期のデータに対してマイニングを行い、初期のマイニング結果33を生成する。

【0043】新規マイニング系は、図1と同様に追加データ発生部21，新規マイニング部22，検証部23および合成部24により構成される。このシステムによると、合成部24の出力がマイニング結果として次回に用いられる。

【0044】例えば毎月1回データが追加されるような場合に月単位で追加データに対してマイニングを行った場合、月毎のマイニング結果にかなりのばらつきが存在すると考えられる。一方、データを追加してからデータベース全体に対してマイニングを行うと、全体を通して頻度の高い規則のみが抽出される。

【0045】従来ではこの両方の規則を抽出するには、追加データに関するマイニングと全体のマイニングの2つのマイニングを行う必要があった。本発明の手法では追加データに対するマイニングを基本とし、全体に対するマイニングを行うことなく全体を通して頻度の高い規則を効率よく求めることが可能となる。

【0046】そこで、以下にデータが連続的に追加される例を説明する。最初のマイニングを行う時刻を0とし、時刻1、2、3、4でそれぞれデータの追加があったとする。時刻0でのデータ件数、および各時刻において追加されるデータの件数はそれぞれ1000件とする。最小サポート値は0．1、すなわち各時刻において追加されるデータの中で100件以上の頻度の規則を抽出するものとする。

【0047】時刻0～4について追加データのマイニングが行われた結果、表14に示すように6種の規則について、各時刻において追加されるデータ内での頻度が得られたと仮定する。

【0048】

表14

時刻	0	1	2	3	4
規則1	200	160	180	150	140
規則2	150	40	30	10	10
規則3	120	120	80	90	120
規則4	100	60	110	70	100
規則5	80	130	120	140	150
規則6	40	50	150	120	90

即ち、各時刻に追加されるデータのみについてマイニングを行うと、結果として頻度が100以上の規則が得られる。すなわち、表14で下線部分がマイニング結果として出力される。

【0049】次に、各時刻においてデータを追加した後、全体に関してマイニングを行った場合について説明する。各規則の頻度は、その時刻までの頻度の累積値となり、表15のようになる。

【0050】

表15

時刻	0	1	2	3	4
規則1	200	360	540	690	830
規則2	150	190	220	230	240
規則3	120	240	320	410	530
規則4	100	160	270	340	440
規則5	80	210	330	470	620
規則6	40	90	240	360	450

この場合は、時刻0で100以上、時刻1で200以上、時刻2で300以上、時刻3で400以上、時刻4で500以上の規則がマイニング結果として出力される。すなわち、表15で下線部分が結果として出力される。

【0051】本発明の手法は、図7に示すように合成部において、各時刻のマイニング結果として、規則、開始時刻、累積頻度の3つの情報を以下の手順により生成し、保存および再利用するものとする。

【0052】先ず、規則が累積マイニング結果33に含まれているかが判定される（S51）。この判定がYESであれば、即ち過去のマイニング結果に含まれている規則ならば、過去のマイニング結果の累積頻度に現在時刻の追加データの頻度を加えて規則を出力し（S54）、開始時刻はそのままとする（S55）。

【0053】ステップ51での判定がNOであれば、即ち過去のマイニング結果に含まれていない規則であり、現在時刻の追加データの頻度が最小サポート値より高ければ、累積頻度を現在の時刻の追加データの頻度として規則を出力し（S52）、開始時刻を現在時刻とする（S53）。

【0054】この手法を上記の例に適用すると、各時刻でのマイニングの出力は下表16のようになる。

【0055】

表16

	規則	開始時刻	累積頻度
時刻0	規則1	0	200
	規則2	0	150
	規則3	0	120
	規則4	0	100
時刻1	規則1	0	200+60=360
	規則2	0	150+40=190
	規則3	0	120+120=240
	規則4	0	100+60=160
	規則5	1	130
時刻2	規則1	0	360+180=540
	規則2	0	190+30=220
	規則3	0	240+80=320
	規則4	0	160+110=270
	規則5	1	130+120=250
	規則6	2	150
時刻3	規則1	0	540+150=690
	規則2	0	220+10=230
	規則3	0	320+90=410
	規則4	0	270+70=340
	規則5	1	250+140=390
	規則6	2	150+120=270
時刻4	規則1	0	690+140=830
	規則2	0	230+10=240
	規則3	0	410+120=530

規則4	0
規則5	1
規則6	2

このようにすると、ある時刻において追加されるデータの中で一度でも最小サポート値以上の頻度のある規則は、その後ずっとマイニング結果として出力されることとなる。すなわち、任意の時刻においてデータベース全体についてマイニングして得られる結果はすべてこのリストの中に含まれる。

【0056】なお、本手法ではマイニング結果がデータを追加する度に増加するため、マイニングの実行時間が増加する可能性がある。その改良として、累積頻度の比率が一定以下になったとき出力する規則を除去する方法も考えられる。例えば、累積頻度の比率が0．05以下になったら規則を結果より除去するとすると、時刻4で規則2が除去される。このような判断は、開始時刻と各時刻に追加されるトランザクション数を保持すれば容易に計算できる。

【0057】(第2の実施形態)第1の実施形態ではデータベースが追加される場合について述べたが、過去1年

340+100=440
390+150=540
270+90=360

間というようにデータベース内に格納するデータの期間を一定とする使い方をされる場合がある。この場合は新しいデータを追加する度に、期間をはずれたデータを除去する必要があり、マイニング結果の保持に関しても除去を考慮する必要がある。

【0058】以下に、本発明の第2の実施形態に従った周期的なインクリメントマイニングシステムを図8を参照して説明する。

【0059】図8の構成によると、図6のシステムに時刻別マイニング結果41が付加されている。このシステムを第1の実施形態で用いた例と同じデータで説明する。すなわち、時刻0-5における規則1-6の出現頻度を表14と同じものを用いる。

【0060】ここで、周期は3、すなわち過去3回のデータを保持するものとする。周期を3としたときのデータベース全体のマイニング結果を表17に示す。

【0061】

表17

時刻	0	1	2	3	4
規則1	200	360	540	490	470
規則2	150	190	220	80	50
規則3	120	240	320	290	290
規則4	100	160	270	240	280
規則5	80	210	330	390	410
規則6	40	90	240	320	360

この場合は、時刻0で頻度が100以上、時刻1で200以上、時刻3以降では300以上の規則がマイニング結果として出力される。すなわち、上記の表17で下線部分が結果として出力される。

【0062】以下では周期3において、追加部分のマイニング結果と過去のマイニング結果より全体のマイニング結果を求める手法について図9のフローチャートを参照して説明する。

【0063】時刻2までは第1の実施形態と同一であり、時刻3のときに時刻0のデータを除去して時刻3のデータを追加し、時刻4では時刻1のデータを削除して時刻4のデータを追加する。マイニング結果としては、第1の実施形態と同様にデータベース全体について成り立つ規則に関して規則内容、開始時刻、累積頻度を特定する情報を保持するのに加え、各時刻における追加データに関するマイニング結果41、すなわちデータの追加時点で出力される規則の追加データにおける出現頻度を保持するものとする。各時刻における手順は図9のフローチャートに示されるように行う。

【0064】先ず、規則が累積マイニング結果33に含まれているかが判定される（S61）。この判定がYESであれば、即ち、規則が過去のマイニング結果に含まれている規則であれば、開始時刻が1周期前以前かが判定される（S62）。この判定がYESであれば、累積頻度が直前の累積頻度－削除時の頻度＋現在時刻の頻度で算出される（S63）。即ち、一定期間の累積マイニング結果は累積マイニング結果を追加データによって検証して得られるマイニング結果から削除すべき期間のマイニング結果を減じ、追加のマイニング結果を合成することによって求められる。開始時刻は1周期前＋1とされる（S64）。

【0065】ステップS61での判定がYESであり、ステップS62での判定がNOであれば、累積頻度が直前の累積頻度＋現在時刻の頻度によって求められ（S65）、開始時刻はそのままの値とされる（S66）。

【0066】ステップS61の判定がNOであれば、過去のマイニング結果に含まれていない規則において、現存時刻の追加データにおける頻度が最小サポート値より高ければ、累積頻度を現在時刻の追加データにおける頻度として規則を出力し（S67）、開始時刻を現在時刻とする（S68）。

【0067】上記の手順に従った周期3とした場合の各時刻におけるマイニング結果を表18に示す。

【0068】

表18

	規則	開始時刻	累積頻度
時刻0	規則1	0	200
	規則2	0	150
	規則3	0	120
	規則4	0	100
時刻1	規則1	0	200+60=360
	規則2	0	150+40=190
	規則3	0	120+120=240
	規則4	0	100+60=160
	規則5	1	130
時刻2	規則1	0	360+180=540
	規則2	0	190+30=220
	規則3	0	240+80=320
	規則4	0	160+110=270
	規則5	1	130+120=250
	規則6	2	150
時刻3	規則1	1	540+150-200=490
	規則2	1	220+10-150=80
	規則3	1	320+90-120=290
	規則4	1	270+70-100=240
	規則5	1	250+140=390
	規則6	2	150+120=270
時刻4	規則1	2	490+140-160=470
	規則2	2	80+10-40=50
	規則3	2	290+120-120=290
	規則4	2	240+100-60=280
	規則5	2	390+150-130=410
	規則6	2	270+90=360

明らかに、本方式において出力されるマイニング結果は、データベース全体について行ったマイニング結果を含む。また、第1の実施形態と同様に、頻度が一定以下になった規則をマイニング結果から削除することも容易である。

【0069】上述のように本発明によると、データの追加・削除があったときに過去のデータベースにアクセスすることなく、過去のマイニング結果を追加データについて検証して得たマイニング結果と追加データに関するマイニング結果とを合成することにより全体のマイニングを行う。

【0070】

【発明の効果】本発明によれば、データベースにデータが追加されるとき、データベース全体をマイニングすることなく、追加されるデータのマイニングと追加される以前のデータベースのマイニング結果を合成することによりデータベース全体のマイニングが可能となり、大規模データのマイニングを効率よく実行するために有効である。

【0071】また、データの追加時にもっとも古い時刻のデータを削除するような周期的なデータベースにおいても同様に過去のマイニング結果を利用してデータベース全体のマイニングが可能となり、大規模データのマイニングを効率よく実行するために有効である。

【図面の簡単な説明】

【図1】本発明の一実施形態に従ったインクリメンタルマイニングシステムのブロック図。

【図2】本発明の過去のマイニング結果を得るためのインクリメンタルマイニング方法を説明するフローチャート。

【図3】第1の実施形態に従った新規マイニング結果を得るためのインクリメンタルマイニング方法を説明するフローチャート。

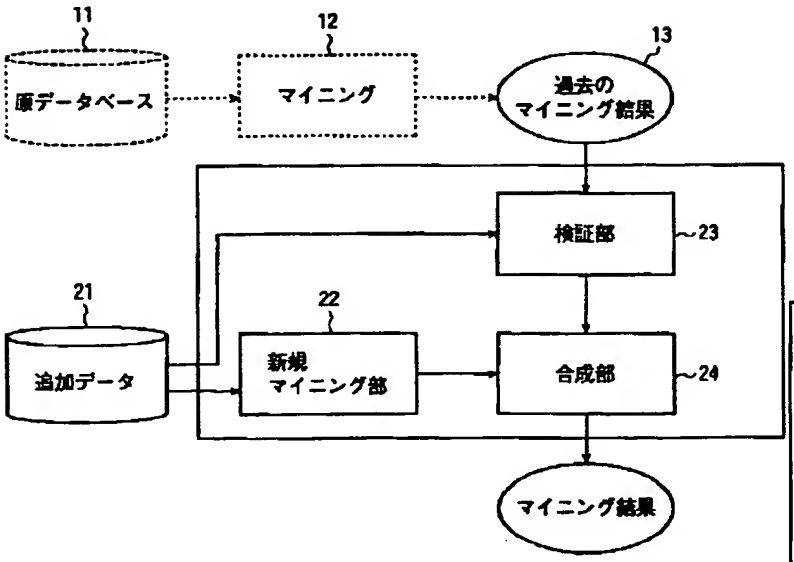
【図4】第1の実施形態の新規マイニングで使用する検証部を説明するフローチャート。

【図5】第1の実施形態の新規マイニングで使用する合

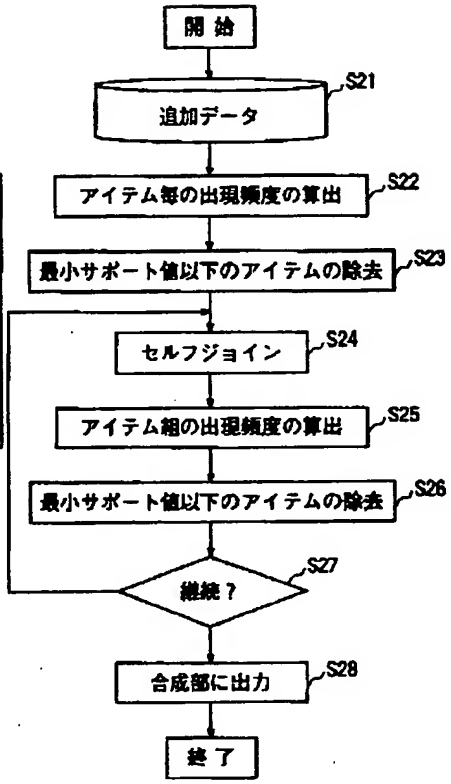
成部を説明するフローチャート。
【図6】初期マイニング結果を用いるインクリメンタルマイニングシステムのブロック図。
【図7】図6のマイニングシステムにおける合成部を説明するフローチャート。
【図8】本発明の第2の実施形態に従ったインクリメンタルマイニングシステムのブロック図。
【図9】図8のマイニングシステムにおける合成部を説明するフローチャート。
【符号の説明】

11…原データベース
12…過去マイニング部
13…過去のマイニング結果
21…追加データ部
22…新規マイニング部
23…検証部
24…合成部
31…初期データベース
32…初期マイニング部

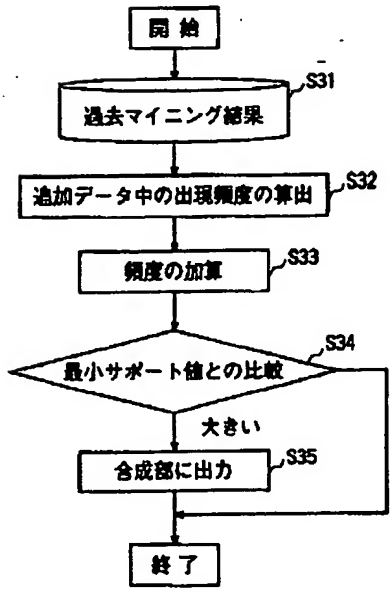
【図1】



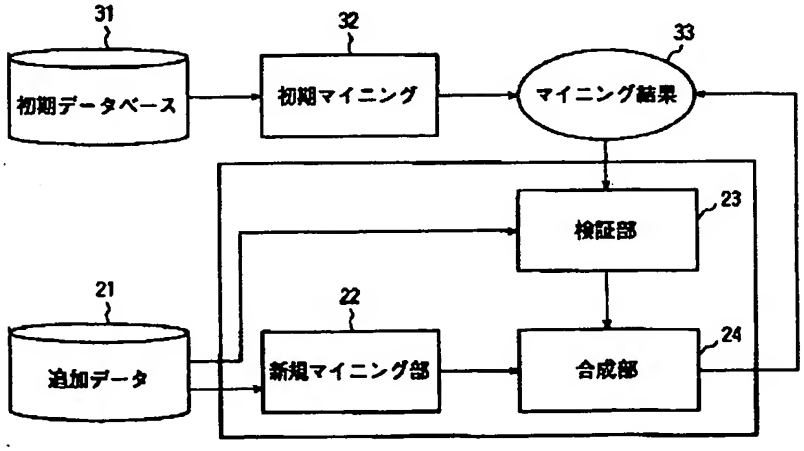
【図3】



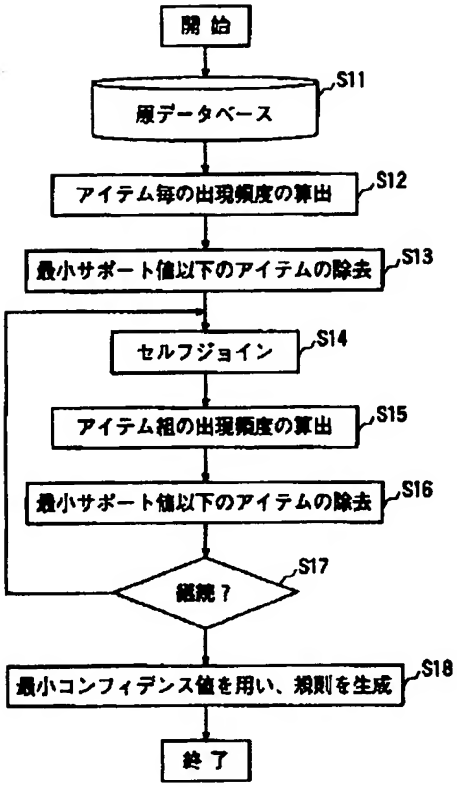
【図4】



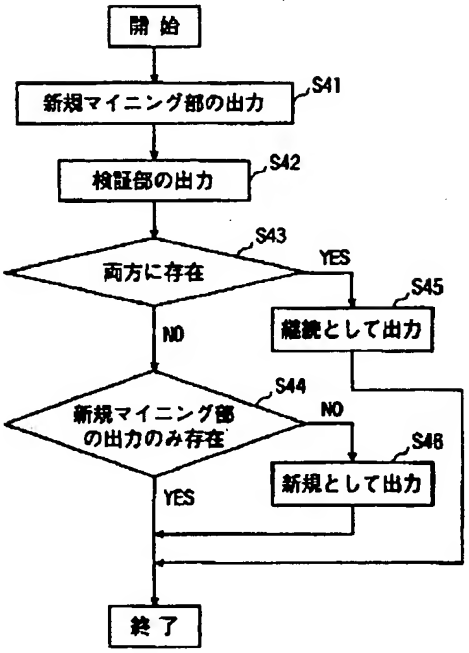
【図6】



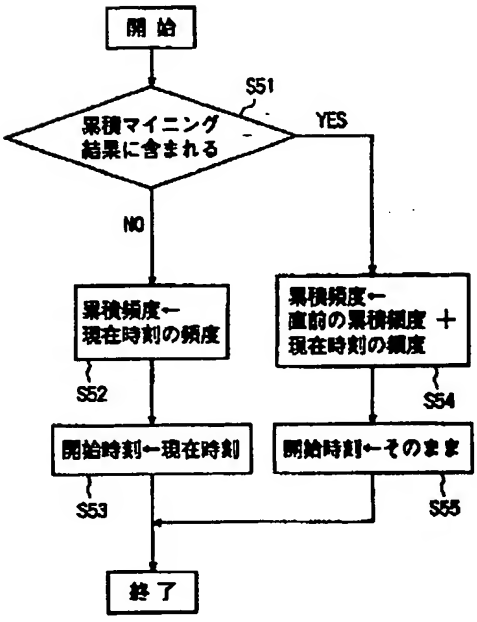
【図2】


開始
S11
履データベース
アイテム毎の出現頻度の算出
S12
最小サポート値以下のアイテムの除去
S13
セルフジョイン
S14
アイテム組の出現頻度の算出
S15
最小サポート値以下のアイテムの除去
S16
継続？
S17
最小コンフィデンス値を用い、規則を生成
S18
終了

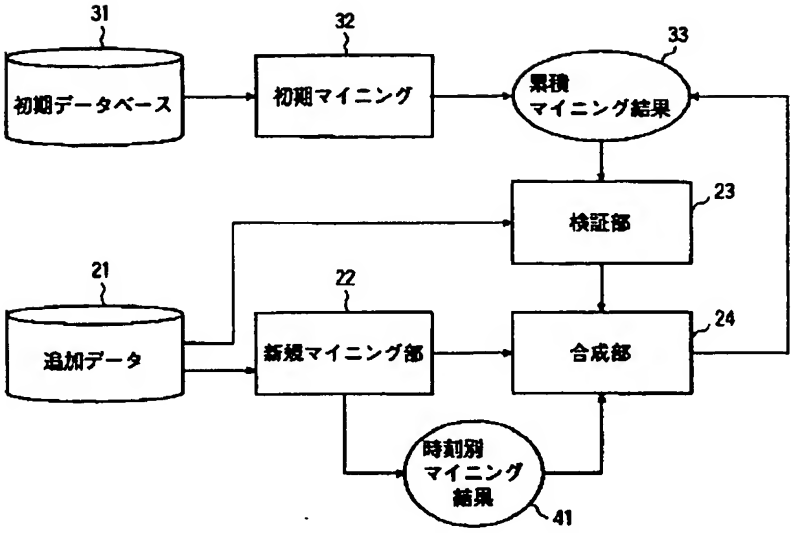
【図5】


開始
新規マイニング部の出力
S41
検証部の出力
S42
両方に存在
S43
YES
S45
継続として出力
NO
新規マイニング部の出力のみ存在
S44
NO
S46
新規として出力
YES
終了

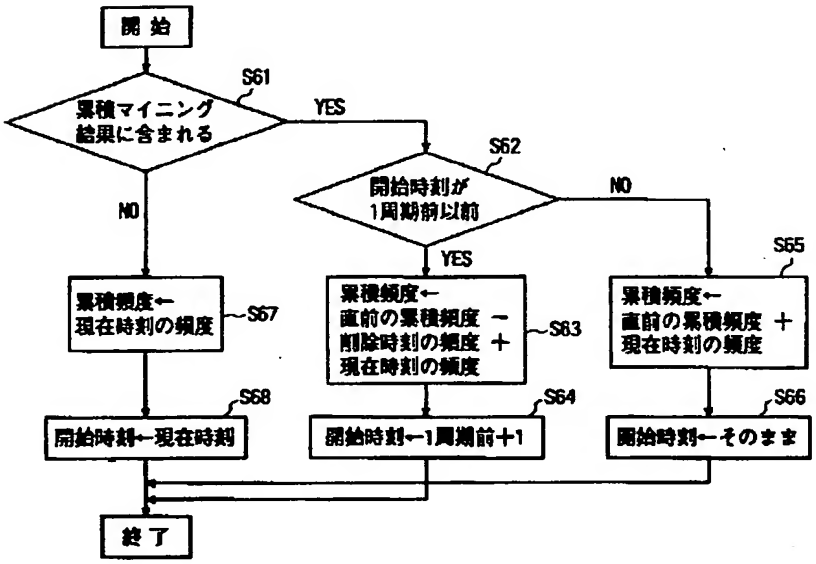
【図7】


開始
累積マイニング結果に含まれる
S51
YES
NO
累積頻度←現在時刻の頻度
S52
累積頻度←直前の累積頻度＋現在時刻の頻度
S54
開始時刻←現在時刻
S53
開始時刻←そのまま
S55
終了

【図8】



【図9】



フロントページの続き

(72)発明者　仲瀬　明彦
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝研究開発センター内

(72)発明者　久保田　和人
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝研究開発センター内

Fターム(参考)　5B075 PR04
5B082 GA03